知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

国内外公共政策文本分析中主题模型应用研究进展

龙艺璇 [1,2] 伊惠芳 [1,2]
[1] 中国科学院文献情报中心 北京 100190
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100049

摘要：[**目的 / 意义**] 梳理主题模型在公共政策文本中的国内外应用现状有助于学习已有研究成果，为未来发展提供理论与实践支持。[**方法 / 过程**] 采用文献计量分析法从时间趋势、机构分布、期刊分布等角度进行量化分析，详细归纳阐述应用现状；其次，通过关键词共现识别国内外主要研究方向并展开对比分析，总结主题模型应用在公共政策文本中存在的问题并提出未来展望。[**结果 / 结论**] 公共政策文本分析中主题模型的应用整体呈增长态势，前景广阔。国内外研究起步时间相当，但国内研究在研究范围、研究深度、合作方式、研究方法等方面均需提升。此外，未来发展存在主题模型自身方法适用性问题和研究内容粒度问题，需进一步结合公共政策文本特征改进主题模型并细化研究力度。

关键词：主题模型　公共政策　文本分析　LDA
分类号：G250
DOI：10.13266/j.issn.2095-5472.2020.029

引用格式：龙艺璇 , 伊惠芳 . 国内外公共政策文本分析中主题模型应用研究进展 [J/OL]. 知识管理论坛 , 2020, 5(5): 305-316[引用日期]. http://www.kmf.ac.cn/p/225/.

1 引言

公共政策是指国家机关及其他权威机构在一定时期为实现特定目标所采取的政治行为或规定的行为准则，它包括法律、规划、措施、方法、办法、条例、通知、意见等 [1]，具有价值取向特定、主客体明确、权威性、强制性等基本特征 [2]。政策文本的内容解读可以在一定程度上帮助了解一个国家的执政理念和战略规划，如今科学技术日新月异、国际环境复杂多变，各国政策颁布层出不穷，政策文本量与日俱增，数据密集型科学的到来给公共政策内容分析带来了新的挑战。高效解读大量公共政策文本内容，可以为公共政策领域众多研究提供有力的基础支持。

作者简介：龙艺璇 (ORCID：0000-0002-5395-4049)，博士研究生，E-mail: longyixuan@mail.las.ac.cn；伊惠芳 (ORCID:0000-0003-0094-7993)，博士研究生。

收稿日期：2020-07-07　　发表日期：2020-10-23　　本文责任编辑：刘远颖

诞生于 20 世纪 90 年代的文本挖掘技术提供了大规模文本内容分析的新契机，如 J. Li 等采用多种文本挖掘算法设计商业政策文档流程分析框架[3]；L. Prior 等将文本挖掘与语义网分析相结合，揭示英国卫生政策构成基本要素[4]；J. Y. Lee 等运用文本挖掘方法分析研究中美在双边贸易和“一带一路”等重大外交政策上的差距[5]；K. Misook 等采用大数据分析软件 Textom 对韩国体育政策进行文本分析并可视化[6]。随着研究不断深入，有学者意识到用传统的文本挖掘方法开展公共政策文本分析得到的结果可解释性较差，无法满足细粒度的信息需求[7]，因此亟需适应大数据文本且深入语义层面的文本挖掘技术改善这一现状。

1999 年，T. Hofmann 首次提出主题模型 PLSA（Probabilistic Latent Semantic Analysis），实现了对文本中深层潜在语义进行挖掘[8]。主题模型的诞生为主题挖掘提供了更多的可能性，改善了基于传统文献计量方法（如词频分析、共词分析[9]、引文分析[10-11]）开展主题挖掘时存在的引文时滞、共词高低词频等不足，众多研究人员根据特定任务目的和情境对主题模型进行改进。如目前适用性较广的隐含狄利克雷分布模型（Latent Dirichlet Allocation，LDA）[12]，能够捕获文档库中主题动态变化的动态主题模型（Dynamic Topic Models, DTM）[13]、将作者信息融入主题模型从而建立“作者 - 主题”关联的作者主题模型（Author-Topic Model，ATM）[14-15] 等。目前，主题模型已经广泛应用到文本聚类[16]、主题演化[17] 等众多研究中。有学者开始尝试使用主题模型挖掘公共政策文本内容，这主要取决于主题模型的特点能够与公共政策文本的特性相吻合，适用性主要表现在以下 3 个方面：①主题模型适用于大数据非结构化文本，与公共政策大规模文本量和非结构化特性相吻合；②主题模型可以实现文本语义降维，挖掘潜在语义关系，因此适用于公共政策文本的高维特性；③主题模型可以较为准确高效地识别大规模文档中的多主题，这与公共政策文本的多主题特性相契合。可以预料，主题模型实现公共政策文本内容的梳理与解读是可行的，并在未来会有更长足的发展。

主题模型在公共政策文本分析中的应用仍处于起步阶段，目前尚未有学者系统梳理相关研究方法与研究内容，学界对主题模型在公共政策文本应用研究缺乏系统全面的认知，不利于学习和借鉴已有的研究成果和研究方法，也限制了主题模型在公共政策文本分析中的优化与扩展应用。基于此，笔者将研究视角定位于主题模型应用在公共政策文本分析中的相关研究，采用文献计量方法，借助统计分析和关键词共现，重点关注主题模型是如何应用在公共政策领域以及利用主题模型解读公共政策文本后可以解决公共政策领域的哪些问题，总结归纳出国内外目前发展存在的局限性，并指出未来可能的发展方向。

2 主题模型在公共政策文本中应用量化分析

2.1 数据来源

笔者选择 Web of Science 核心合集和 CNKI 学术期刊全文数据库作为数据来源数据库。考虑到主题模型目前有很多改进和衍生算法，如 PLSA[8]、LDA[12] 等，为保证检索结果尽可能检全，笔者依据算法名称充分扩充检索词。同时，因部分缩写存在一定歧义，人工对全部检索结果依据题目和摘要进行筛选。此外，笔者重点关注的是将主题模型应用在公共政策文本中的研究，而不是应用在公共政策领域中的研究，因此最终筛选结果中所有文献的主题建模对象应为各类公共政策文本，而非论文、专利等科学文献。通过人工筛选得知，检索结果中大部分文献主要内容集中在使用主题模型分析某一研究领域研究进展并附带提出该领域相关政策建议，此类研究虽在主题中也涉及公共政策，但均以期刊论文或专利文本为主题建模对象，与本文关注的公共政策文本不符，因此也被剔除，这也是最终

知识管理论坛，2020(5):305–316
DOI：10.13266/j.issn.2095–5472.2020.029

人工筛选结果与检索结果数量差异较大的原因。具体检索过程及检索结果见表 1。需说明的是，本文的检索式只能保证检索到在主题中明确提出“policy”或“政策”的目标文献，然而有些公共政策文本如通知、意见、措施等并不会带有“policy”或“政策”字眼，本文的检索过程在一定程度上有可能会忽略掉部分目标文献。笔者认为，即使对公共政策文本进行主题建模的目标文献研究对象为“通知”“意见”“措施”等，但绝大多数学者会在主题中提及“政策”或“policy”，因此还是采用了表 1 中的检索式，并结合人工筛选保障检索结果的准确性。

表 1 检索过程及检索结果

数据类型	检索日期	来源数据库	检索式	检索结果	人工筛选最终结果
国际数据	2020.6.1	Web of Science核心合集	TS=(“topic model*” OR LDA OR “Latent Dirichlet Allocation” OR PLSA OR PLSI OR “probabilistic latent semantic analysis” OR “Probabilistic Latent Semantic Indexing”)AND (Policy OR Policies) AND 文献类型: (Article) 不限时间跨度	157篇	23篇
国内数据	2020.6.1	CNKI学术期刊全文数据库	主题 (精确) = (主题模型 OR 主题建模 OR LDA模型 OR 潜在狄利克雷分布 OR 隐含狄利克雷分布 OR 潜在狄利克雷分配 OR 隐含狄利克雷分配 OR PLSA OR PLSI OR 概率潜在语义索引 OR 概率隐含语义索引 OR 概率潜在语义分析 OR 概率隐含语义分析) AND (政策) 不限时间跨度	152篇	19篇

2.2 时间趋势分析

科研文献数量随时间的变化可以在一定程度上反映相关研究发展状况 [18]。由于 2020 年非完整自然年，因此不考虑在内，时间分布统计结果见图 1，虽然在 1999 年主题模型就已诞生，但直到 2015 年才有学者尝试将主题模型应用在公共政策文本中。整体来看目前研究数量不多，国际与国内研究起步时间相当，近几年均呈明显上升趋势。从增长速度来看，国际数据增长略快于国内数据增长。

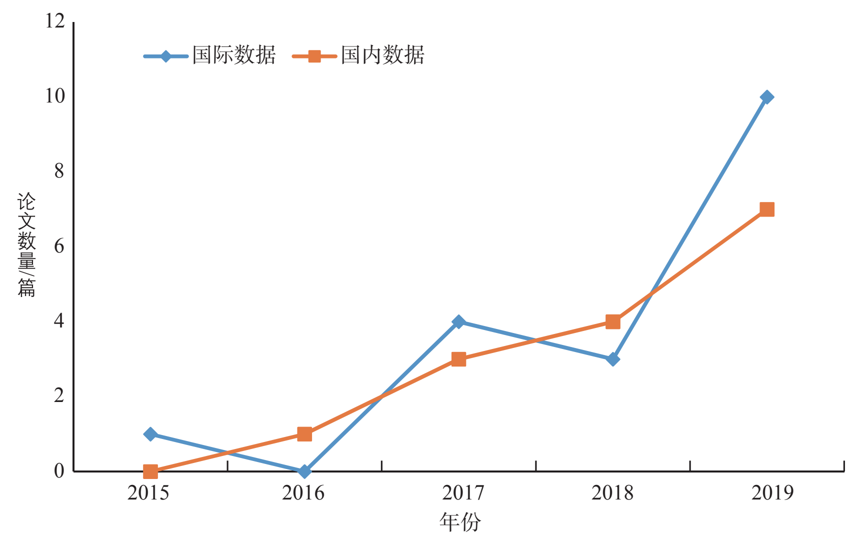


图 1 主题模型在公共政策文本中应用的发文时间分布

2.3 发文机构分布

使用全计数法统计发文机构结果见图 2 和图 3，从中可以看出国内外研究机构均较为分散。此外，通过对作者合著现象统计分析发现，国际上发表的 23 篇相关文献中，有 11 篇为多机构合作，而国内发表的 19 篇相关文献中，只有 3 篇为多机构合作，因此可以得出国际研究更倾向于多个机构之间共同合作，而国内更倾向于单一机构内的学者展开合作。从国际数据中的机构国别来看，美国发表的文献居多，占所有国际数据的 1/3 以上。从机构形式来看，国际数据和国内数据均是以高校发文为主，研究所发文为辅。

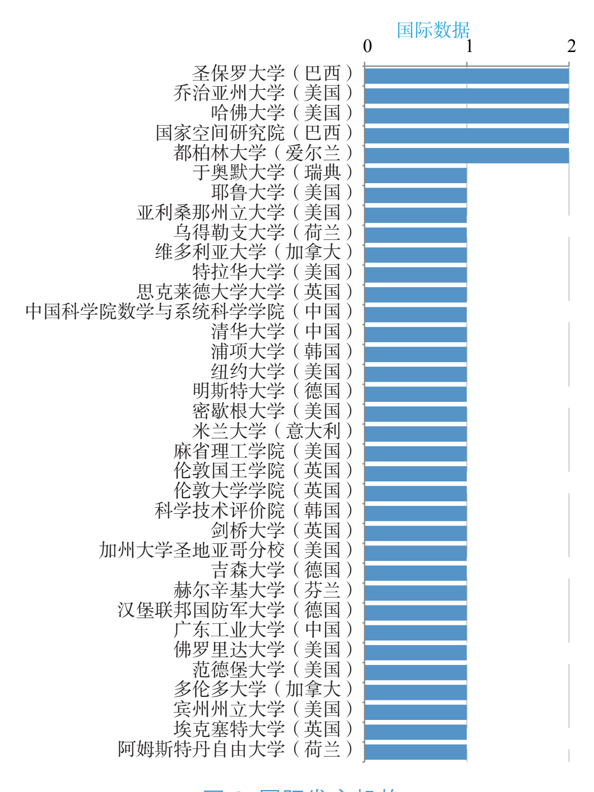


图 2 国际发文机构

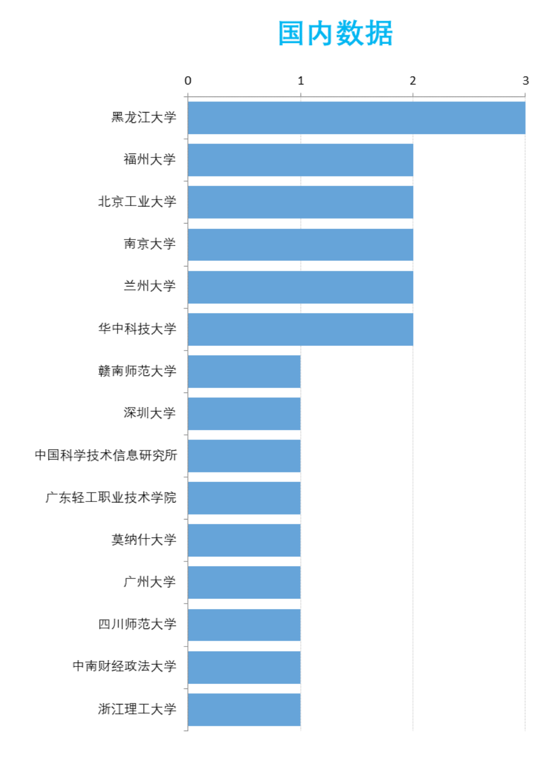


图 3 国内发文机构

2.4 期刊分布

发文期刊统计结果见表 2。从发文期刊领域来看，发现国际数据中发文期刊主要集中在政策研究领域的期刊，而国内数据主要集中在情报学领域的期刊。此外，笔者还发现，国内学者在国际上发表的相关论文更倾向于领域特色非常明显的专业期刊。

3 主题模型在公共政策文本中研究方法分析

众多学者根据研究目标和研究文本实际情况对主题模型 PLSA 进行改进，逐渐诞生了 LDA、DTM、ATM、TOT 等一系列适应不同研究需求的主题模型。鉴于公共政策文本存在非结构化、高维、多主题等特性，为进一步分析目前主题模型具体方法在公共政策这一特殊文本中的应用，笔者根据检索结果对国内和国际数据中主题模型具体使用算法进行统计，结果见图 4 和图 5。

通过对比两图可知，国内在公共政策文本分析中主题模型使用较为单一，绝大部分学者采用目前最主流的 LDA 主题建模方法开展相关研究，只有极少数学者根据实际研究情况采用考虑了时间因素的主题时间模型（TOT）[19]。

知识管理论坛，2020(5):305–316
DOI：10.13266/j.issn.2095–5472.2020.029

表 2 主题模型在公共政策文本中应用的发文期刊

国际期刊名称	国际发文数量/篇	国内期刊名称	国内发文数量/篇
POLITICAL ANALYSIS	2	现代情报	2
JOURNAL OF RURAL STUDIES	1	数据分析与知识发现	2
AMERICAN JOURNAL OF POLITICAL SCIENCE	1	情报杂志	2
APPLIED ECONOMICS LETTERS	1	情报探索	2
CLIMATE POLICY	1	信息资源管理学报	1
ECOLOGICAL INFORMATICS	1	全球科技经济瞭望	1
ENERGY RESEARCH & SOCIAL SCIENCE	1	情报理论与实践	1
ENVIRONMENTAL DEVELOPMENT	1	情报科学	1
EUROPEAN JOURNAL OF POLITICAL ECONOMY	1	兰州大学学报（社会科学版）	1
GEORGE WASHINGTON LAW REVIEW	1	广州大学学报（自然科学版）	1
GOVERNANCE-AN INTERNATIONAL JOURNAL OF POLICY ADMINISTRATION AND INSTITUTIONS	1	福州大学学报（哲学社会科学版）	1
GOVERNMENT INFORMATION QUARTERLY	1		
INTERNATIONAL JOURNAL OF BEHAVIORAL NUTRITION AND PHYSICAL ACTIVITY	1		
INTERNATIONAL PUBLIC MANAGEMENT JOURNAL	1		
JOURNAL OF CLEANER PRODUCTION	1		
NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE	1		
PARTY POLITICS	1		
POLICY SCIENCES	1		
POLITICAL SCIENCE RESEARCH AND METHODS	1		
POLITICS AND GOVERNANCE	1		
TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	1		
TRANSPORT POLICY	1		

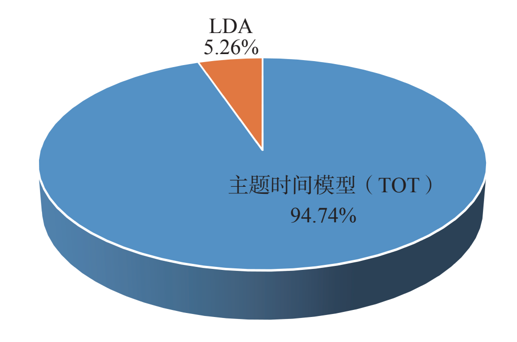


图 4 主题模型在公共政策文本分析中研究方法分布（国内数据）

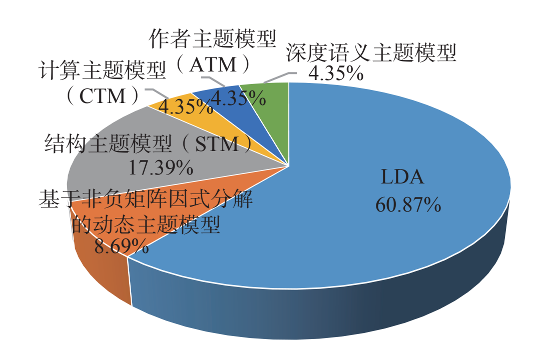


图 5 主题模型在公共政策文本分析中研究方法分布（国际数据）



知识管理论坛
2020 年 第 5 期(总第 29 期)

国际上公共政策文本分析中主题模型方法使用更加多样化，虽然 LDA 依然占据绝对优势，但有部分学者积极尝试使用结构主题模型（STM）、计算主题模型（CTM）、作者主题模型（ATM）以及基于非负矩阵因式分解的动态主题模型，此外，还有学者使用了 Leximancer（一种文本分析软件）开展政策文本主题建模，该分析软件中内嵌基于深度学习的深度语义主题模型[20]。

4 主题模型在公共政策文本中研究内容分析

为更加直观分析主题模型在公共政策文本中的应用方向，笔者借助 Vosviewer 软件采用关键词构建共现网络，并采用归纳研究法进一步总结。

4.1 国内研究内容分析

首先将国内数据导入 Vosviewer 分析软件，对关键词进行手工筛选后，最小聚类大小设为 30，得到国内主题模型在公共政策文本中应用方向，见图 6。

红色关键词代表方向 1，根据“专题数据库”“政策分析系统”“政策文本管理”“LDA”“政策结构”等关键词，结合国内相关文献具体内容，将该主题研究方向总结为公共政策文本组织与管理研究。李少博[21]采用 LDA 主题模型对科技政策文本进行建模，构建基于主题的科技政策分析系统；王倩倩[22]采用 LDA 模型对科技政策检索用户的信息与检索记录进行主题建模，开发科技政策领域个性化语义检索系统；张涛等[23]通过引入政策词表和对 LDA 模型进行加权的方式，提出一种新型政策文本聚类方法；刘雨农等[24]采用 LDA 主题模型对政策文本开展主题分类，并结合词频统计归类，提出人文社科专题数据库主题选择框架，为人文社科专题数据库建设提供支持。

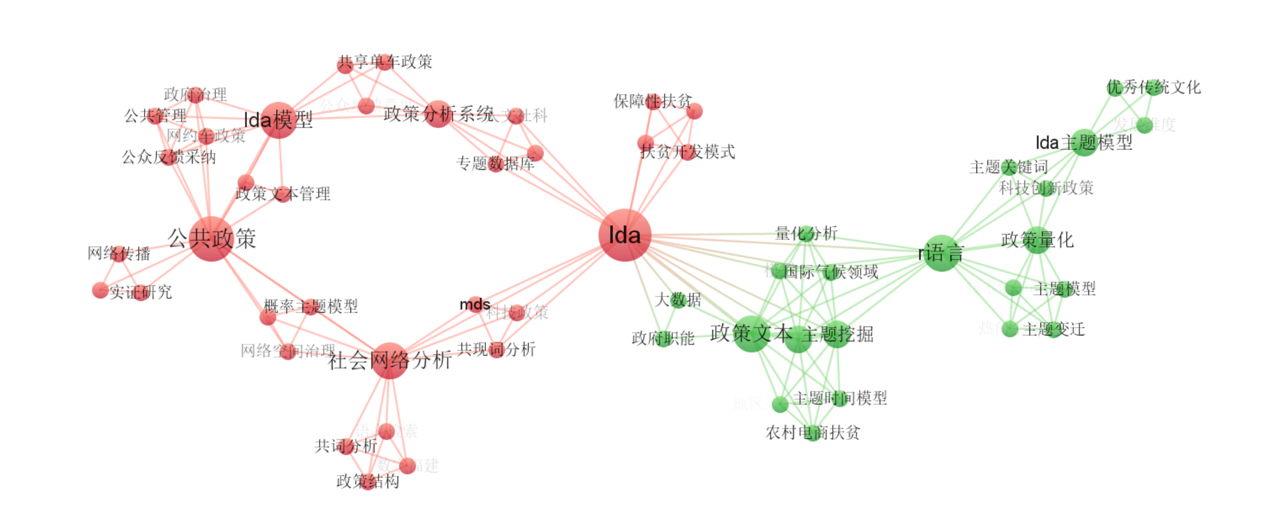


图 6 主题模型在公共政策文本分析中研究内容分布（国内数据）

绿色关键词代表方向 2，“政策文本”“主题时间模型”“主题变迁”“lda”“量化分析”等关键词均表达出内容随时间变化的含义，结合文献内容，归纳该方向下的主要研究内容为公共政策主题演化研究。余传明等[19]运用融入抽取词时间戳的 TOT 主题时间模型，得出农村电商扶贫政策的时间 - 主题概率分布以及主题 - 词汇概率分布，分析农村电商扶贫政策内容演化情况；杨慧等[7]以气候相关政策文本为研究对象，基于 R 语言改进 LDA 主题模型，

知识管理论坛，2020(5):305–316
DOI：10.13266/j.issn.2095–5472.2020.029

开展政策文本主题内容及主题强度演化趋势分析；张永安等[25]收集国家、北京、中关村三级技术创新政策，运用LDA主题模型识别主题，为技术创新政策的完善提出相关建议；郎玫等[26]开展不同地区政策主题并与中央政府职能匹配研究，得出地方政府存在行政职能弱化和职能供给结构性不足等问题。

4.2 国际研究内容分析

将国际数据导入Vosviewer分析软件，对关键词进行手工筛选后，最小聚类大小设为30，得到国际主题模型在公共政策文本中应用方向，如图7所示：

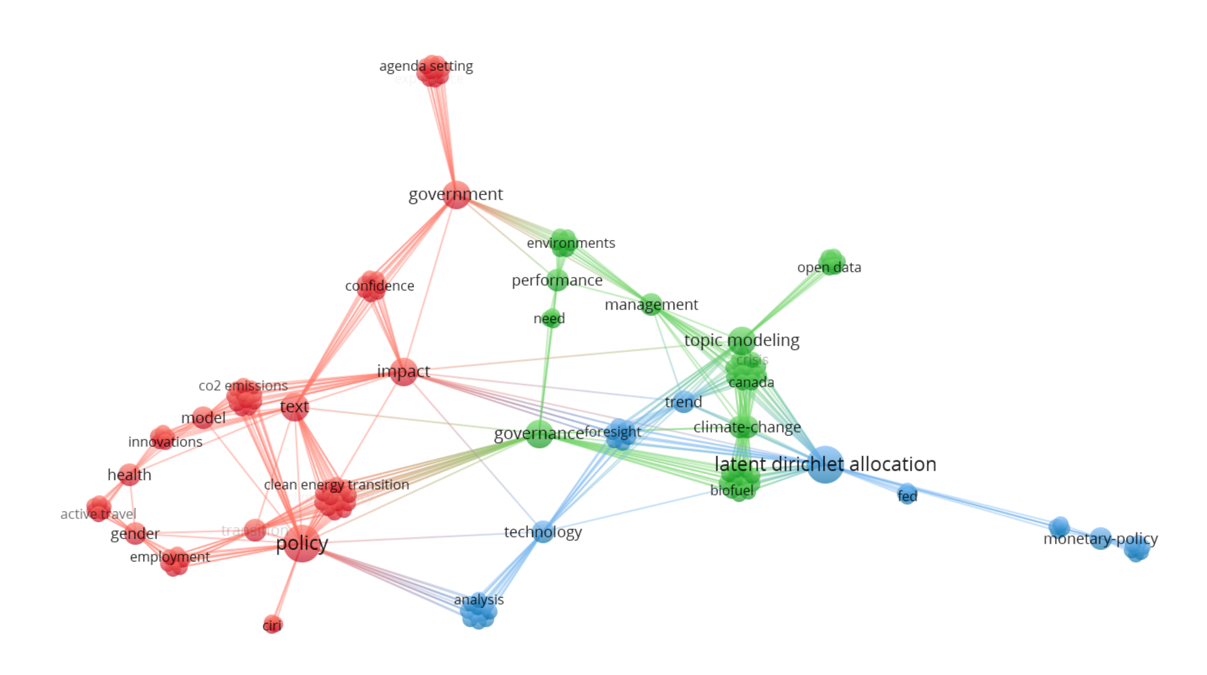


图7 主题模型在公共政策文本分析中研究内容分布（国际数据）

蓝色关键词代表方向1，根据“latent dirichlet allocation”“policy”“trend”“foresight”等主题词，结合文献内容将该方向归纳为公共政策主题演化研究。该研究方向与国内研究方向2类似，均是利用主题模型分析公共政策内容随时间的变化。如2019年A. Mark等[27]以生态领域为例，采用LDA模型、HDP（Hierarchical Dirichlet Process）和TF-IDF分析，对美国政府文件进行主题分析。Q. Wen等[28]收集桥梁管理（BM）相关的政策法规作为数据集，采用作者-主题模型（ATM）文本挖掘的方法识别政策中的关键主题。

绿色关键词代表方向2，根据“topic modeling”“management”“performance”“need”等关键词，结合文献具体内容归纳该方向为公共政策文本组织与管理研究。该方向主要研究内容与国内研究方向1相似，均是利用主题模型分析实现大规模公共政策内容高效的组织管理，以期实现公共政策内容的妥善保存和便捷利用。如C. Lucas等[29]采用结构化主题模型实现政策文本的自动化分析，便于随时把握政策最新进展；J. B. Ruhl等[30]将研究对象集中在法律文件，利用LDA主题模型实现法律文件实质性的主题分类，并且比较了传统方法与主题建模方法的优缺点。

红色关键词代表方向3，根据“text”“impact”“policy”“topic model”等关键词，结合文献具体内容将该方向内容归纳为利用主题模型开展政策影响研究。该研究方向目前国内鲜有学者涉及。该主题既包括政策

实施带来的影响，也包括其他因素对政策产生影响。如 H. S. Du 等 [31] 采用 LDA 模型对中国各省环保部门官网的环境政策数据开展文本挖掘，检验绿色投资的空间特征以及政治、经济和环境因素的溢出效应；A. Ceron 等 [32] 采用结构主题模型来分析 74 份议案、1 439 份演讲和 9 份大会宣言中包含的内容，以评估派系动议或个别演讲是否对政党宣言中的内容产生了影响。

此外，通过逐篇回顾国际文献内容，发现个别英文文献难以划分到具体的研究方向中，通过阅读文献归纳其主要内容涉及不同区域政策内容比较、项目评价等。H. Ale 等 [33] 应用结构主题模型（Structural Topic Modelling，STM）分析 147 个国家有关全球气候治理研究的政策，比较发展中国家和发达国家关于全球气候治理关注的关键主题；K. Isoaho 等 [34] 对欧盟 5 000 多个政策文件进行主题建模分析，来证实能源联盟（Energy Union）项目的政策优先级。

4.3 国际与国内研究内容对比分析

笔者总结国际与国内主题模型在公共政策文本中应用主要存在以下几点不同：

首先，从研究内容范围来看，国际学者将主题模型应用在公共政策文本中的范围更广，尝试利用新方法解决更多的传统问题。国内学者研究的主题主要集中在公共政策主题演化研究和公共政策文本组织与管理研究，而国外研究除了以上两个研究方向，还尝试将主题模型应用在公共政策影响、不同区域政策内容比较、项目评价等相关研究。

其次，从研究内容时间来看，近两年国际学者研究主要集中在利用主题模型解决公共政策影响评估和公共政策内容演化问题，较少关注到公共政策文本组织与管理研究，而国内学者自开始尝试将主题模型应用到政策文本中，应用方向过多局限在解决公共政策文本组织与管理问题以及政策内容演化问题，鲜有学者尝试拓展。

最后，从研究内容重视度来看，国际学者最重视的是利用主题模型开展公共政策影响相关研究，试图利用大规模文本分析弥补以往政策影响难以定量化评估的缺陷。而国内学者将主要科研精力放在了公共政策内容演化与公共政策文本组织管理研究，倾向于使用自动化的方法应对公共政策文本量剧增带来的公共政策内容精炼与政策文本管理问题，偏重于公共政策文本本身而忽略了与其他文本相结合。

笔者尝试从更深层角度分析国际与国内研究内容出现差异的原因，主要有以下 3 点：

首先，国内研究虽重视程度日益增加，但缺乏系统综述。在本文之前，国内尚未有综述性文章总结国外目前主题模型在政策文本中的应用现状，这不利于国内学者学习和借鉴国外相关研究方向和研究成果，因此才会造成国内研究方向较为局限的现状。

其次，国内学者学科背景较为单一。国内主要是图书情报领域研究学者应用主题模型分析政策文本，并发表在图情领域期刊。而国际包含了政策领域、图情领域、资源环境领域等多领域学者，多发表在政策研究期刊。学科的单一限制了思维的扩展，国内图情领域的学者更希望主题模型在分析政策文本过程中可以解决图情领域传统问题，而国际上不同领域的学者面临的问题不同，因此更愿意尝试从不同的角度应用主题模型，这也进一步解释了国内研究内容近几年一直没有太多应用方向上的创新而国际研究方向逐渐多样。

最后，与国际相比国内机构间合作少。合作更容易碰撞出思想的火花。与国内研究相比，国际研究机构间合作更加紧密，更容易产生新思路与新方法，产生更多新思路与新方法。因此，在将主题模型应用在政策文本分析时，国际学者关注的不仅仅是政策文本自身，而是尝试与其他文本相结合，探索政策文本与其他文本之间的关系。而目前国内合作范围较为狭窄，不利于国内学者进一步拓宽研究视野，这也在一定程度上解释了国内学者始终将研究定位于政策文本本身而国际学者在多源文本对比中开拓了新研究方向。

知识管理论坛，2020(5):305-316
DOI：10.13266/j.issn.2095-5472.2020.029

5 问题与展望

5.1 现有研究存在的问题

笔者认为国内外公共政策文本分析中主题模型应用局限性主要表现研究方法和研究内容两个层面。

首先，在研究方法上，目前应用最广泛的LDA主题模型本身就存在一定的缺陷，如最优主题数量一般依据经验设定 [35] 或者使用计算复杂度较高困惑度来确定 [12]，前者强烈依赖人工经验，后者则需要较高的计算时间成本；主题由主题词表征，语义揭示性不强，可解释性不够 [36]；只能表征文档 - 主题、主题 - 主题词纵向关系，无法利用主题模型揭示主题和主题之间的横向关系等 [37]。LDA 虽然适用于大规模文本分析，但其自有的缺陷将严重阻碍在政策文本中的广泛应用。此外，目前已有学者已经意识到 LDA 主题模型的缺陷，并尝试使用改进过的主题模型（如 Time Dynamic Topic Models、ATM、TOT、STM）等分析政策文本，但目前使用的主题模型改进多是基于论文或专利文本，鲜有学者根据公共政策文本的具体特征进一步改进主题模型，主题的可解释性仍有很大提升空间。

此外，在研究内容上，相比于论文的摘要、关键词等结构化表示，政策文本结构性较差，现有研究主要针对政策文本的全部内容，而公共政策包含政策目标、政策工具、政策效果、政策主体、政策对象等诸多要素，使用主题模型识别出的政策主题只能在整体层面表示政策的主要内容及变化，无法深入细致到某一类政策要素，研究缺乏针对性。

除了以上两点共性问题，国内研究还存在研究思维固化、合作缺乏、领域单一等局限。首先，虽然国内学者紧跟国际步伐将主题模型应用在政策文本分析中，但研究中心始终定位于公共政策文本自身，忽略了与其他文本相结合的新思路；其次，机构间合作较少，不利于碰撞出新的思维火花；最后，参与研究人员学科背景较为单一，限制了思维的扩展。

5.2 未来展望

主题模型在政策文本中的应用仅仅是一个尝试性的开始，具有广阔的发展空间。针对上述目前研究存在的共性问题，笔者认为可以从以下两点进行改进：

首先，针对研究方法上的局限性，需要研究人员从公共政策文本特性出发，在借鉴以往对主题模型改进思路与方法基础上，尝试对主题模型进行改进。政策文本除了具备文本词项高维性、主题复杂性和长文本特征之外，结构性相比于传统分析文本更差，同时也不具备技术词、专业术语等代表性词语，不同种类的政策文本表达方式也相对多样化，以上特点均要求应用在政策文本中的主题模型应具备更高的可解释性和可理解性。

针对研究内容上的局限性，为进一步满足科研人员和决策者对政策内容的分析需求，未来主题模型在公共政策文本中的应用应更加精细化，考虑聚焦于政策文本中的单一要素，实现细粒度信息需求的满足，如政策工具作为保障政策目标顺利实现的重要手段，政策工具的演化分析对于政策制定者和科学研究者来说意义重大，目前绝大多数研究均采用内容分析法，需要依靠人工编码，亟需实现政策工具要素的自动抽取与内容分析。此外，随着文本挖掘技术的不断改进和主题模型可解释性的不断提升，可考虑进一步扩展研究范围，尝试应用主题模型解决更多政策领域存在的研究问题。

此外，针对国内研究存在的问题，除了需要改进以上两点，还需在重视程度、研究范围、研究深度、合作方式等方面做出努力。首先，增加主题模型在文本分析领域的重视程度，密切跟踪国际最新应用动态，总结国际经验，争取密切跟进国际研究步伐；其次，积极扩展研究思路，考虑政策文本与其他文本相结合，在解决传统问题的基础上，争取有新发现；最后，加强机构间和国际间合作，除了加强同领域机构间的合作，还需要加强跨领域合作，融合不同学科的思路，尝试解决不同学科的问题，还

可以积极与国际其他研究机构合作，进一步融入国际科研圈，共同探索主题模型在公共政策文本分析中的更多可能性。

6 结语

本研究通过梳理国内外公共政策文本分析中主题模型的应用研究现状，得出目前国内和国际研究者都在积极尝试在公共政策文本分析中使用主题模型，但在合作方式、期刊分布领域、研究方向等方面仍存在较大差异。首先，在合作方式方面，国际上发表的相关文献更倾向于多个机构共同合作，而国内更倾向于单一机构内的学者展开合作；其次，在发表期刊分布领域方面，国际研究发文期刊主要集中在政策研究领域的期刊，而国内研究主要集中发表在情报学领域的期刊；最后，在研究方向方面，国际学者关注研究方向更加广泛，积极尝试使用新方法解决多种研究问题，且随时间推移，近几年国外应用方向更加分散，而国内学者研究方向相对固化，研究思维不够发散，忽略了政策文本与其他文本的结合。目前，国内外公共政策文本分析中主题模型应用在研究方法和研究内容上均存在一定局限性，但毫无疑问未来大规模公共政策文本的分析将更加依赖于主题模型等深入语义的文本挖掘算法，具有广阔的发展空间。未来需要有针对性地提升主题模型对公共政策文本的适用性，拓展研究深度与广度，提高分析效率和分析结果的可解释性，为政策研究提供有力支撑。

参考文献：

[1] 陈振明 . 政策科学——公共政策分析导论 [M]. 北京 : 中国人民大学出版社 , 2003: 19.
[2] 苏竣 . 公共科技政策导论 [M]. 北京 : 科学出版社 , 2014: 8-9.
[3] LI J, WANG H J, ZHANG Z, et al. A policy-based process mining framework: mining business policy texts for discovering process models[J]. Information systems and e-business management, 2010, 8(2): 169-188.
[4] PRIOR L, HUGHES D, PECKHAM S. The discursive turn in policy analysis and the validation of policy stories[J]. Journal of social policy, 2012, 41(2): 271-289.
[5] LEE J Y, LEE J. A text mining analysis of US-Chinese leaders on trade policy[J]. Journal of international logistics and trade, 2019, 17(3): 67-76.
[6] MISOOK K. Trends of sports policy through the analysis of big data text-mining: with a focus on the inauguration of the MCST minister[J]. The Korean journal of sport, 2019, 17(2): 519-529.
[7] 杨慧 , 杨建林 . 融合 LDA 模型的政策文本量化分析——基于国际气候领域的实证 [J]. 现代情报 , 2016, 36(5): 71-81.
[8] HOFMANN T. Probabilistic latent semantic analysis[C]// Fifteenth conference on uncertainty in artificial intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1999: 289-296.
[9] LIU L Q, MEI S Y. Visualizing the GVC research: a co-occurrence network based bibliometric analysis[J]. Scientometrics, 2016, 109(2): 1-25.
[10] DEREK J S P. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.
[11] GARFIELD E. Citation indexes for science: a new dimension in documentation through association of ideas[J]. Science, 1964, 144(3619): 649.
[12] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. The journal of machine learning research, 2003, 3(3): 993-1022.
[13] BLEI D M, LAFFERTY J D. Dynamic topic models [C]// Proceedings of the 23rd international conference on machine learning. New York: ACM Press, 2006: 113-120.
[14] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]// Proceedings of the 20th conference on uncertainty in artificial intelligence. Arlington: AUAI Press, 2004: 487-494.
[15] ROSEN-ZVI M, CHEMUDUGUNTA C, GRIFITHS T, et al. Learning author-topic models from text corpora[J]. ACM transactions on information systems (TOIS), 2010, 28(1): 4-42.
[16] 曲靖野 , 陈震 , 郑彦宁 . 基于主题模型的科技报告文档聚类方法研究 [J]. 图书情报工作 , 2008, 62(4): 113-120.
[17] 王丽 , 沈湘 . 文本预处理后的 LDA 模型主题发现与技术演进研究 [J]. 农业图书情报 , 2019, 31(4): 19-28.
[18] 曹树金 , 吴育冰 , 韦景竹 , 等 . 知识图谱研究的脉络 , 流派与趋势——基于 SSCI 与 CSSCI 期刊论文的计量与可视化 [J]. 中国图书馆学报 , 2015, 41(5): 16-34.

知识管理论坛，2020(5):305–316
DOI：10.13266/j.issn.2095–5472.2020.029

[19] 余传明，郭亚静，龚雨田，等．基于主题时间模型的农村电商扶贫政策演化及地区差异分析 [J]. 数据分析与知识发现，2018, 19(7): 34-45.
[20] HAYNES E, GREEN J, GARSIDE R, et al. Gender and active travel: a qualitative data synthesis informed by machine learning[J]. International journal of behavioral nutrition and physical activity, 2019, 16(1): 135-146.
[21] 李少博．基于主题的科技政策分析系统设计与实现 [D]. 石家庄：石家庄铁道大学，2016.
[22] 王倩倩．科技政策领域的个性化语义检索系统研究 [D]. 石家庄：石家庄铁道大学，2016.
[23] 张涛，马海群．一种基于 LDA 主题模型的政策文本聚类方法研究 [J]. 数据分析与知识发现，2018, 21(9): 59-64.
[24] 刘雨农，吴柯烨，权昭瑄．人文社科专题数据库建设的主题选择研究 [J]. 现代情报，2019, 39(12): 11-18.
[25] 张永安，马昱．基于 R 语言的区域技术创新政策量化分析 [J]. 情报杂志，2017, 36(3): 113-118.
[26] 郎玫．大数据视野下中央与地方政府职能演变中的匹配度研究——基于甘肃省 14 市 (州) 政策文本主题模型 (LDA)[J]. 情报杂志，2018, 37(9): 78-85.
[27] MARK A, CHRISTOPHER B, JESSE A. Documents as data: a content analysis and topic modeling approach for analyzing responses to ecological disturbances[J]. Ecological informatics, 2019, 51: 82-95.
[28] WEN Q, QIANG M, XIA B Q, et al. Discovering regulatory concerns on bridge management: an author-topic model based approach[J]. Transport policy, 2019, 75: 161-170.
[29] LUCAS C, NIELSEN R A, ROBERTS M E, et al. Computer-assisted text analysis for comparative politics[J]. Political analysis, 2015, 23(2): 254-277.
[30] RUHL J B, NAY J, GILLIGAN J M. Topic modeling the president: conventional and computational methods[J]. George Washington law review, 2018, 86(5): 1243-1315.
[31] DU H S, ZHAN B Q, XU J H, et al. The influencing mechanism of multi-factors on green investments: a hybrid analysis[J]. Journal of cleaner production, 2019, 239(1): 1-12.
[32] CERON A, GREENE Z. Verba volant, scripta manent? Intra-party politics, party conferences, and issue salience in France[J]. Party politics, 2019, 25(5): 701-711.
[33] HSU A, BRANDT J, WIDERBERG O, et al. Exploring links between national climate strategies and non-state and subnational climate action in nationally determined contributions (NDCs)[J]. Climate policy, 2019, 19(6): 443-457.
[34] ISOAHO K, MOILANEN F, TOIKKA A. A big data view of the European Energy Union: shifting from a floating signifier to an active driver of decarbonisation?[J]. Politics and governance, 2019, 7(1): 28-44.
[35] 伊惠芳，吴红，马永新，等．基于 LDA 和战略坐标的专机技术主题分析——以石墨烯领域为例 [J]. 情报杂志，2018, 37(5): 97-102.
[36] 王丽，沈湘．文本预处理后的 LDA 模型主题发现与技术演进研究 [J]. 农业图书情报，2019, 31(4):19-28.
[37] 刘自强，许海云，岳丽欣，等．基于 Chunk-LDAvis 的核心技术主题识别方法研究 [J]. 图书情报工作，2019, 63(9): 73-84.

作者贡献说明：

龙艺璇：提出研究命题，设计研究方案，进行数据处理与分析，撰写论文内容；

伊惠芳：收集与分析研究数据，修订论文内容。

知识管理论坛
2020 年 第 5 期(总第 29 期)
www.kmf.ac.cn

Application of Topic Models in the Analysis of Public Policy: A Review of the Research Status in Domestic and Foreign

Long Yixuan[1,2] Yi Huifang[1,2]
[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049

Abstract: [Purpose/significance] This paper comprehensively summarizes the application of topic models in public policy texts, which helps researchers learn from existing research results and provides theoretical and practical support for future development. **[Method/process]** This paper used bibliometric analysis to study from the perspectives of time trend, organization distribution, periodical distribution, etc., and summarized the application status in detail. Secondly, the LDA topic model was used to identify the main international and domestic research directions and conducted a comparative analysis. Finally, this paper summarized the problems in the application and proposed future prospects. **[Result/conclusion]** The application of topic models in the analysis of public policy texts is on the rise overall and has broad prospects. The starting time of domestic and foreign research is equivalent, but domestic research needs to be improved in terms of research scope, research depth, cooperation methods, and research methods. In addition, in the future development, there are problems with the applicability of the topic model's own methods and the granularity of research content. It is necessary to further combine the characteristics of public policy texts to improve the topic model and refine research efforts.

Keywords: topic model pubic policy text analysis LDA